様式 C-19

科学研究費補助金研究成果報告書

平成 23　年 6　月　3 日現在

機関番号：14301
研究種目：若手研究(B)
研究期間：2009 年　　～　2010 年
課題番号：21700162
研究課題名（和文）　事象知識の体系化－事象の因果関係・信憑性解明のための基盤技術－

研究課題名（英文）　Systematization of Knowledge about Events -Fundamental Techniques for Analyzing Causal Relationships and Credibility of Events-

研究代表者
浅野 泰仁　　(ASANO YASUHITO)
京都大学・情報学研究科・特定准教授
研究者番号：20361157

研究成果の概要（和文）：因果関係の解析と情報の信憑性判定支援に役立つ事象知識の体系化を行った．具体的には，以下の(1)-(3)を行った．(1)事象知識が時間とともに伝播していく仕組みを解析する時間グラフパターンマイニング手法の提案 (2)事象知識のモデルと抽出手法の提案 (3)事物間の関係の強さと関係を成り立たせている事物を求める手法の提案と Wikipedia 画像信憑性判定支援への応用

研究成果の概要（英文）：We have systemized knowledge about events useful for analyzing causal relationships and helping judgments of information credibility. The summary of our work is the following (1)-(3): (1) Time graph pattern mining methods for analyzing propagation of eventual knowledge. (2) A model of eventual knowledge and a method of extracting eventual knowledge. (3) A method for computing the strength of the relationship between two objects and obtaining objects elucidating the relationship. We have also applied ideas used in this method to helping a user to judge the credibility of an image on Wikipedia.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,300,000	390,000	1,690,000
2010 年度	1,100,000	330,000	1,430,000
年度			
年度			
年度			
総　計	2,400,000	720,000	3,120,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：ウェブ，マイニング，関係に関する知識，事象知識，時間グラフパターン

１．研究開始当初の背景

現在は情報爆発の時代であり，世界で起きた様々な事象に関するニュースや記事が Web 上に氾濫している．たとえば各マスコミのサイトには，テレビや新聞で報道されるニュース(マスコミ系ニュース)が載っているし，ブログサイト等にはマスコミが報道しないような事象に関するニュースや記事が多く見られる．これらの事象の因果関係や，事象に関する情報の信憑性を解明することが，今後の社会において重要となるという認識は広まっている．特に，マスコミ系ニュースのみならず，Web 一般のページや Wikipedia には，因果関係・信憑性解明に役立つ重要な情報が多く含まれる．一例としては，メーカーが報道しないような，製品を実際に使用したユーザーによる不具合報告や，マスコミが報道する以前から中国の偽塩の有毒性につ

いて指摘していた記事などが挙げられる．
しかしながら，それらの事象に関する記述から信憑性・因果関係解明のための情報を整理してデータベース化する手法も提案されていないため，せっかくの膨大な情報が知識として活かされていないのが現状である．
既存研究としては， Web からある程度の情報を整理してデータベースに登録する研究や，新聞記事から文章の 5W1H 表現を抽出する手法，文章同士の因果関係を抽出する手法，マスコミ系ニュースサイトの記事をクラスタリングする手法は存在する．また，事物間の関係の強さをモデル化する手法はあるが，その関係の強さを説明する事物を発見する手法は存在しなかった．事象の因果関係や信憑性を解析するために Web から事象知識を抽出し知識として利用できる形でデータベース化する手法も存在しない．

２．研究の目的

本研究では，事象情報の因果関係・信憑性解明のための，事象知識の体系化の手法を確立する研究を行う．
事象知識を自動的に収集し，それらが形成するグラフ構造のパターンを調査し，有用な知識を発見する手法を提案する．特に，事物間の関係に着目し，その強さをモデル化すると同時に，その関係を説明する事物を発見する手法を構築し，この手法によって体系化された関係の知識を DB に格納する．また，この手法を事象知識の信憑性検証に応用する．

３．研究の方法

上記の目的を達成するために，以下の(1)-(3)を行う．
(1) 収集したグラフ構造のパターンを調査する．特に，事象情報では時間が重要となるため，特徴的な時系列情報を持つものを調査する．例として，Web 上で短期間に盛り上がった話題や，Amazon の書籍が時系列を追って次々に出版され，その購入データから推薦ネットワークが生じていくパターンを調査する．グラフパターンマイニングの既存研究を基に，これらを解析して有用な知識を発見する手法を提案する．

(2) Web ページなどから事象知識を自動的に抽出する，すなわち各事象の 5W1H(Who, What, When, Where, Why, How)キーワードを発見する手法を提案する．この情報を RDB に格納する．

(3) 事物間の関係をモデル化する手法，それも関係の強さと関係を成り立たせている事物両方を同時に求めることのできる手法を提案する．この手法によって体系化された関係の知識を RDB に格納する．また，この手法を，事象知識を大量に含んでいる Wikipedia の画像の信憑性判定支援に応用する．

４．研究成果

上記研究方法で挙げた(1)-(3)それぞれの成果について概説する．

(1) 時間と構造の情報を同時にとらえることのできる“時間グラフパターン”とそのマイニング方法を新たに提案した．時間グラフパターンは web やソーシャルネットワークなどといった特定のネットワークには依存しないため，これまでのグラフパターンと同様多くのネットワークに対して利用できる．
マイニングのために，節点ごとに時間的特徴をその節点のラベルへと変換する．そのために用いる時系列データとしてグラフ中の節点数を用いる手法，検索エンジン上でクエリ数を用いる手法の二つを提案する．
時間グラフパターンが時間と構造的特徴を検索に有効かどうかを実験により確かめた．まずサンプルとなる話題からネットワークを生成し，パターンをマイニングする．次に得られたパターンをマッチングに用いることで，別のネットワークから同一の構造を持つ部分グラフを抽出する．このマッチング結果を解析することにより，対象ネットワークから有用な知見を得る．我々は“web 上の話題の盛り上がり”，および“書籍推薦ネットワーク上での技術トレンド”という二種類の実験を行なった．以下，それぞれについて簡単に説明する．
まず，web 上の話題の盛り上がりは web グラフが時間ともにリンク構造が成長したものと見なすことができる．我々は時間グラフパターンを用いた実験によって，話題の盛り上がり時に重要な役割を果たす web ページの三つの役割を発見し，その役割に応じて web サイトを分類することに成功した．その役割ごとに節点の生成期間や周辺のリンク構造は特有のものとなっており，時間グラフパターンによってとらえられる．そのため，これらの役割は web の時間と構造的特徴を反映していると言える．
次に，オンラインショッピングサイトの amazon のデータから，節点が書籍，辺が書籍間の推薦を表す書籍推薦ネットワークを作成し，この上で技術トレンドについて解析した．まずサンプルとなる技術に関連する書籍からネットワークを生成し，パターンをマイニングする．一つの例として“データマイニング” という技術を取り上げ，マイニングされたパターンを用いた書籍のサブカテゴリー分類を行なった．これらのサブカテゴリーは書籍の出版日と技術トレンドの成熟度の

関係を反映したものになっているとともに，特徴的な推薦構造を持っている．これらの特徴を利用して，書籍の検索に応用できることも示す．これらの実験を通じて，提案した時間グラフパターンがさまざまなネットワークに対して有用であることが分かった．これにより，時間グラフパターンを用いれば，時間情報も考慮したこれまでにないネットワーク解析や情報検索が行えることが分かった．

(2) Wikipedia の記事集合から事象に関する知識を抽出するためのモデル及び手法を提案した．さらに，これを用いて同一の事象に関する記述を発見する手法を提案した．発見された同一の事象に関する記述を比較することで，記事の閲覧者は事象について深く理解することが可能となる．また，この手法は情報の不整合を発見し，記事の信憑性を検証するのにも役立つ．同一事象に関して発見された複数の記事の記述の不整合を提示することで，閲覧者は誤った情報を発見することができる．

事象を扱うモデルとして Wikipedia Sentential Event モデル(以下 WSE モデル)を提案し，このモデルを用いて事象に関する情報を獲得，事象情報データベースに格納する．WSE モデルでは句点で区切られた 1 文と 1 事象として扱う．事象を表す要素として，5W1H に関する情報のうち When,Who,Where の要素に着目する．これらの情報はすべての事象について記述されているわけではない．そのためこれらの情報では事象を特定するには不十分であり，さらに動作の主体となる述語，および When,Who,Where 以外の重要名詞をキーワード集合として用いる．

評価実験として安倍晋三の記事を起点に，各記事にリンクしている記事を収集し，手法を適用し，得られた事象に関する知識データベースを作成した．また，その有効性の検証のため，そのデータベースを用いて，同一の事象の判定を行ったところ，適合率 0.86，再現率 0.33 となり，高精度で同一事象に関する記述を発見できることがわかった．

(3)
(i) 距離，連結度，共引用の三つの概念全てに基づき，最大減衰流（Generalized Maximum Flow）を用いて関係を解析する「減衰流モデル」を開発した．二つのオブジェクト(事物) u と v の関係を解析するために，u と v を含むネットワークを構築する．その後，u から v までフローを送り，v に届いた最大となるフローの値で関係の強さを表す．さらに，フローが大量に流れたパスを求めることで，関係に寄与したオブジェクトを発見することもできる．

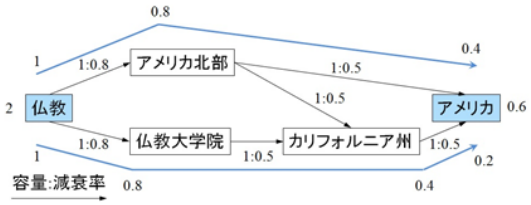

図 1．減衰流モデル

図 1 は減衰流モデルの例である．各辺に容量 1 と 1 より小さい減衰率が設定される．容量とは辺に流れるフローの上限値である．減衰率とは辺の始点から出たフローが終点に到達できる割合である．例えば，「仏教」から「アメリカ北部」へ 1 のフローが流されたときには，「アメリカ北部」に到達するフローはその 0.8 倍の 0.8 となる．したがって，パス（仏教，アメリカ北部，アメリカ）と（仏教，仏教大学院，カリフォルニア州，アメリカ）に沿って，終点「アメリカ」に到達できる最大フローは 0.6 となる．減衰流モデルでは，長いパスに沿って送られるフローが短いパスと比べて比較的小さくなりやすいから，パスの長さを測ることができる．また，フローがほぼ独立したパスに流されるため，近似的にオブジェクト間の連結度を測ることもできる．

減衰流では，フローが始点から終点へ送られるため，始点から終点に向かう方向と反対向きの枝がほとんど使われない．しかし，お互いに向きが異なる枝からなるパスで表現される共引用関係を測るため，反対向きの枝も利用する必要がある．そのために，任意の枝にその枝と方向が異なる枝を追加し，二重化ネットワークを構築する．二重化ネットワークにおいて，最大フローを求めることにより，共引用関係を測ることができるようになる．

減衰流モデルを評価するために，Wikipedia を利用した実験を行った．Wikipedia では，ページをオブジェクトと見なし，ページ間のリンクをオブジェクト間の明示的関係と見なすことができる．Wikipedia のリンク構造により構築した情報ネットワークにおいて，減衰流モデル及び既存手法によってオブジェクト間の関係の強さを求め，減衰流モデルが既存手法と比べてより適切に関係の強さが測れることを確かめた．特に，他の手法と比べ，3 本以上のリンクからなるパスで表現される関係の強さを適切に測ることができることがわかった．

本研究の成果の一つとして，減衰流モデルに関連する技術の特許を出願した．さらに，本研究の成果に関する論文を IEEE の権威ある雑誌「TKDE」に投稿している．

(ii) Wikipedia ネットワークにおいて，オブジェクト s と t の関係を理解するために，その関係を表す重要な独立したパスを抽出する手法を開発した．例えば，図 2 は「石油」と「アメリカ」の関係を説明する 4 本の独立したパスを描いている．ユーザが左から右へ各パスにあるリンクを理解しながら辿って行けば，各パスの意味を理解することができると考えられる．そのためには，パスを構成する各リンク(u, v)の明示的な関係の意味を，ページ u からページ v へのリンクの周辺テキストを読むことにより理解していけばよいと考えられる．例えば，ユーザが図 2 に示しているスニペットを読めば，「オイルショック」を含むパスの意味，すなわちなぜ「オイルショック」が「石油」と「アメリカ」間の関係にとって重要かを理解できるだろう．

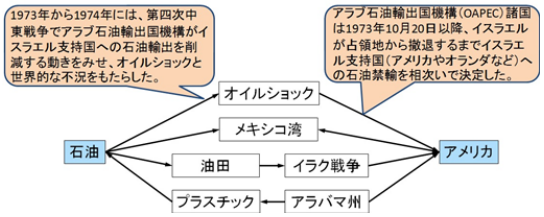

図 2. 関係を説明するオブジェクトとパス

また，この手法で得られた知識を用いて，関係を説明する画像とその周辺テキストをウェブから検索する手法を開発した．

(iii) 縁システムの開発

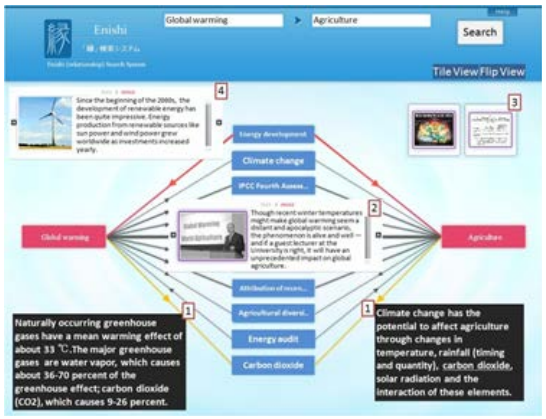

図 3. 縁システム

上記(3)-(i),(ii)を応用し，二つのオブジェクト間の関係に対し，以下の三種類の情報を提供する縁システムを開発した．

(a) 「減衰流モデル」に基づいた手法を利用し，Wikipedia 情報ネットワークから関係に重要な独立したパスを抽出し，表示する．重要なパスが太い線で描かれている．マウスカーソルを枝の上に置けば，図 3 に数字 1 で示しているスニペットのように辺の説明文が表示される．これを読みながらパスの枝を左から右へ辿って行けば，簡単にパスの意味を理解することができる．

(b) 関係に関する知識を含む画像とその周辺テキストを検索する手法を使い，関係全体に関連する知識が含まれている画像とその周辺テキストを検索する．「Flip View」ボタンを押せば，図 3 に数字 2 で示しているウィンドウのように，関係全体に関連する画像と画像周辺テキストを閲覧できる．そして，「Tile View」ボタンを押せば，図 3 に数字 3 で示しているところのように，これらの画像が次々と表示される．これらの画像と画像周辺テキストを読めば，関係をより深く理解することができると考えられる．

(c) 関係全体に関連する画像だけではなく，表示される各パスを説明する画像も検索する．図 3 に数字 4 で示しているウィンドウのように，パスにある辺をクリックすれば，パスに関連する知識を含む画像とその周辺テキストを表示する．関係の構成要素を詳細に調べたい場合には，これらの画像が有用であると考えられる．

縁システムの英語版・日本版ともに http://www.db.soc.i.kyoto-u.ac.jp/enishi/enishi.html からアクセスできるようにしている．また，縁システムに関する論文を Springer 社の権威ある雑誌「Information Retrieval」に投稿中である．

(iv) Wikipedia 画像信憑性への応用

図 4　Wikipedia 画像信憑性判定支援システム

上図は，開発した Wikipedia 画像信憑性判定支援システムの画面である．「日本」という記事に対して，画面上部にある画像がふさわしいかどうかを，その右の棒グラフであらわしている．「関連度」は記事に対するキャプションの適切さを表し，「典型度」は(記事を前提としたときの)キャプションに対する画像の適切さを表している．「ふさわしさ」は関連度と典型度の相乗平均である．数値の範囲は 0 から 100 である．この表示からは，関連度が高く，この画像のキャプションは

「日本」との関係が強く，画像で説明する理由は十分と考えられる．一方，典型度は低く，この画像よりキャプションにふさわしい画像があると考えられる．実際，この画像を見て「北海道が亜寒帯湿潤気候」であるとはわからない．

画面下部の画像集合は，Web を検索して得られた，このキャプションを表すのにふさわしいと考えられる代替画像候補である．実際，右の画像は気候帯を説明しているし，左の画像は北海道が画像にあるほどの大雪が降るような気候，すなわち寒冷で湿潤な気候である証拠となっている．

これらの画像は，「日本」とキャプション中の各リンクの関係を説明する画像集合をクラスタリングし，サイズが比較的大きい各クラスタの中から，典型度が高い画像を選んでいる．典型度は，(3)-(ii)でウェブから取得した画像集合に対して，既存の VisualRank を適用して求めた．

代替画像候補を見ることで，Wikipedia に不足している画像はどのようなものなのか，そしてそれらのこの記事に対するふさわしさを知ることができる．これは，Wikipedia のマルチメディア情報源としての信憑性を確かめるのに大いに有用であると考えられる．Wikipedia の画像よりこれらの画像がふさわしければ，Wikipedia のマルチメディア情報源としての信憑性は(少なくともこの記事に関しては)低いと考えられるからである．

また，評価用に選定した Wikipedia の 15 個の画像(およびキャプションとその画像が載っていた記事)を対象として，評価実験を行った．画像は，人間が問題ない(記事にもキャプションにも適切である)と判断した画像 5 個，記事に適切でない(関連度が低い)事項であると判断した画像 5 個，キャプションに適切でない(典型度が低い)と判断した画像 5 個を集めた．こうして，画像が記事及びキャプションにふさわしいかどうか判定した結果を○×で表した正解集合を作成する．これとシステムの評価を比較した．結果として，15 件中 13 件でシステムの評価と人間の評価が一致し，従って精度は約 86.7%となった．

５．主な発表論文等

（研究代表者，研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

1. Yasuhito Asano, Yuya Miyawaki, and Takao Nishizeki. Efficient Compression of Web Graphs. IEICE Trans. Fundamentals, 査読有，Vol.E92-A, No. 10, pp.2454-2462, 2009.

〔学会発表〕（計 9 件）

1. Xinpeng Zhang, Yasuhito Asano, Masatoshi Yoshikawa. Enishi: Searching Knowledge about Relations by Complementarily Utilizing Wikipedia and the Web. 11th International Conference on Web Information System Engineering(WISE 2010)，査読有，2010.

2. Xinpeng Zhang, Yasuhito Asano, Masatoshi Yoshikawa. Mining and Explaining Relationships in Wikipedia. 21th International Conference on Database and Expert Systems Applications (DEXA2010)，査読有，2010.

3. Taihei Oshino, Yasuhito Asano, Masatoshi Yoshikawa. Time Graph Pattern Mining for Web Analysis and Information Retrieval. 11th International Conference on Web-Age Information Management(WAIM2010)，査読有，2010.

4. Xinpeng Zhang, Yasuhito Asano, Masatoshi Yoshikawa. Analysis of Implicit Relations on Wikipedia: Measuring Strength through Mining Elucidatory Objects. 15th Database Systems for Advanced Applications (DASFAA2010)，査読有，2010.

5. Taihei Oshino, Yasuhito Asano, Masatoshi Yoshikawa. Mining Useful Time Graph Patterns on Extensively Discussed Topics on the Web. 1st International Workshop on Graph Data Management (GDM 2010)，査読有，2010.

6. Katsumi Tanaka, Hiroaki Ohshima, Adam Jatowt, Satoshi Nakamura,Yusuke Yamamoto, Masatoshi Yoshikawa, Qiang Ma, Yasuhito Asano, Kazutoshi Sumiya, Ryong Lee, Daisuke Kitayama, Takayuki Yumoto, Yukiko Kawai, Jianwei Zhang, Shinsuke Nakajima, Yoichi Inagaki. Evaluating Credibility of Web Information. the 4th International Conference on Ubiquitous Information Management and Communication，招待講演, 2010.

7. 俵本 一輝, 川本 淳平, 浅野 泰仁, 吉川 正俊．感情解析のための分布モデルと相互強化型解析手法．第 2 回データ工学と情報マネジメントに関するフォーラム(DEIM 2010)，査読無，2010.

8. 森廣 恭平,張 信鵬,浅野 泰仁,吉川 正俊．Steiner Tree を利用した Wikipedia における

関係の抽出．FIT2009 第 8 回情報科学技術フォーラム．査読無，2009.

9. 張 信鵬,浅野 泰仁,吉川 正俊.減衰流を用いた関係の解析．人工知能学会第 23 回全国大会(JSAI 2009)．査読無，2009.

〔図書〕（計 0　件）

〔産業財産権〕
○出願状況（計 1 件）

名称：関係ネットワーク変換装置，関係ネットワーク変換方法，コンピュータプログラム及び縁検索システム
発明者：浅野泰仁，　張信鵬，吉川正俊
権利者：浅野泰仁，　張信鵬，吉川正俊
種類：発明
番号：2009-142132
出願年月日：2009/6/15
国内外の別：国内

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
浅野 泰仁　　(ASANO YASUHITO)
京都大学・情報学研究科・特定准教授
研究者番号：20361157

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：